DOCUMENT RESUME

ED 400 308 TM 025 679

AUTHOR Cizek, Gregory J.; Fitzgerald, Shawn M.
TITLE A Comparison of Group and Independent Standard Setting.
PUB DATE Apr 96
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Cost Effectiveness; *Group Dynamics; Judges; *Knowledge Level; *Licensing Examinations (Professions); *Physicians; Standards
IDENTIFIERS *Angoff Methods; Experts; Group Process Training; *Standard Setting

ABSTRACT

A group-process approach to standard setting was compared to an independent approach for a medical specialty certification examination. Both approaches used the Angoff (1971) standard-setting method. In the group-process method, reviewers discussed items and their ratings during the rating process; in the independent condition, reviewers provided their ratings in isolation. The effects of having previous exposure to the group-process condition or the independent condition, the effects of knowing other reviewers initial ratings, and the cost effectiveness of the procedures were studied. Participants were 10 subject matter specialists, 5 in each condition. Reviewers in the independent condition made original ratings and then submitted a second rating after they were notified of ratings provided by other reviewers (the "with-information" condition). The results demonstrated fairly large, although nonsignificant, differences in results obtained by group and independent reviewers using the same standard-setting method on identical test content. Although the differences were not statistically significant, a substantial effect on pass-fail decisions was noted. A reviewer's "with information" rating could be fairly well predicted by knowledge of the reviewer's original rating and knowledge of the group mean. Both independent conditions were more economically feasible for the small-panel situation in that they appeared to require a smaller time commitment from participants. (Contains 11 tables and 11 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☑ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
GREGORY J. CIZEK
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 400 308

A Comparison of Group and Independent Standard Setting

April 1996

Gregory J. Cizek
Associate Professor of Educational Research and Measurement
350 Snyder Hall
University of Toledo
Toledo, OH 43606-3390
Phone: 419-530-2611
Email: gcizek@utnet.utoledo.edu

Shawn M. Fitzgerald
Doctoral Candidate
Department of Educational Psychology, Research and Foundations
375 Snyder Hall
University of Toledo
Toledo, OH 43606-3390

Paper presented at the annual meeting of the American Educational Research Association, New York City, NY.

BEST COPY AVAILABLE

TM 025679

A Comparison of Group and Independent Standard Setting

The passing score on an criterion-referenced examination represents the establishment of a standard of performance judged to be acceptable. It is the lowest score that permits the examinee to receive the license or credential. Several methods exist for deriving a passing score; the available methods have been categorized and described in many place, most recently by Jaeger (1989).

The methods comprising what Jaeger has called "test centered continuum models" (1989, p. 493), have been widely used and studied. These methods require standard setting participants (judges, content experts, affected constituencies) to scrutinize items in an examination, usually prior to its administration, and then to provide judgments about the items and/or how examinees would likely respond to the items (e.g., estimate the proportion of minimally competent examinees who will answer the item correctly). For this reason, these methods are also sometimes called "item-based."

One approach to obtaining item reviewers' judgments utilizes a group-process format. In this format, participants are convened in a single location, provided with training in the standard-setting methodology, and directed to provide their ratings for each item in the test. The group-process format is often preferred because, predictably, item reviewers do not produce identical ratings and the group-process format provides a means of discussing differences in perspective, resolving the gross differences in ratings, and promoting reasonableness in the ratings.

Many researchers agree that this reduction of variability is desirable (Jaeger, 1988; Meskauskas, 1986). However, it is common that an extensive portion of a group's meeting

time is devoted to discussions about individual test items, debate, and, when applicable, to consensus-reaching regarding the ultimate rating for each test item.

Norcini, Lipner, Langdon, & Strecker (1987) summarized two of the problems with the group process, including: the tediousness of the task of reviewing individual items and reaching consensus ratings (especially when a large number of items is involved); and, the expense of empaneling a sufficiently large group in one location for, perhaps, several days. Another frequently encountered problem is simply arriving at a single block of time for reviewing items and making the requisite judgments. Hambleton (1978, p. 282) specifically addresses the problem of time resource availability as one of the four primary considerations in selecting a standard-setting methodology.

In addition to the need for research to suggest alternatives for addressing the problems created through use of the group-process format in standard-setting studies, research is needed to examine the effect on standards when such alternative strategies are tried. Fitzpatrick (1989) reviewed literature related to social influences in standard setting. She suggests the following conclusions derived from the literature that may help frame the way interaction is incorporated into standard setting studies:

1. When participants are initially disposed to favor one position over another, discussion of the issue or exposure to the another position will tend to polarize their opinions, with subjective judgments more susceptible to polarization than objective judgments and the polarizing effect of discussion greater than that of exposure;
2. Exposure to an extreme group norm or mean opinion position induces more polarization than simple exposure to a distribution of opinion positions; and
3. Several strategies are known to mediate polarization, including private recording of

judgments, carefully structuring the discussion among participants, and reducing the subjectivity of the judgments they are asked to make.

Fitzpatrick suggests that "We must ask whether it is desirable that the decisions that [item reviewers] make be affected by interpersonal comparisons, by cognitive learning through the exchange of information, or by both types of processes" (1989, p. 321). Mills and Barr have also noted that, "issues of training, group interaction, independent ratings vs. discussion all affect the methods, but little is available in either discussion or guidelines concerning these and other implementation issues" (1983, p. 2). Curry (1987) summarized the state of affairs aptly: "Almost all of these authors [on standard setting] acknowledge that the expert group process will have significant impact on the validity of the outcome, few have examined the dynamics involved" (p. 1).

Objectives

This research examined a group-process approach and an independent approach to standard setting using the Angoff (1971) method on a medical specialty certification examination. For the purposes of this study, a group-process approach and an independent approach were defined as follows: In the traditional group-process method, reviewers discussed items and their ratings during the rating process; in the independent condition reviewers provided their ratings in isolation. The research addressed three objectives: 1) to determine whether item reviewers using the Angoff method produce different ratings as a result of exposure to the group-process condition or an independent condition; 2) to investigate the effect of knowledge of other item reviewers' initial ratings on a subsequent rating of the same items; 3) to conduct a cost-benefit evaluation of the group versus

independent standard setting procedures.

Methods/Data Source

For this study, participants were an existing panel of 10 subject matter experts charged with recommending a passing score for a medical specialty examination. To study differences in group and individual standard setting, participants were randomly assigned to two conditions (n=5 independent; n=5 group-process). It should be noted that, although a sample of 10 participants may be sufficient for operational standard setting, the smaller subsamples used in this study were used for exploratory research purposes only; samples this small would not be recommended in practice.

Prior to a standard setting meeting, all participants were sent materials through the mail to introduce them to the procedures to be used. Subsequently, a whole-group meeting including participants in both groups, was conducted. Together, the groups generated practice Angoff ratings for a sample of nonoperational test items drawn from a recent test form in the medical specialty; practice items were chosen to be representative of items found in the upcoming, operational test form and covered a representative range of difficulty, discrimination, and format. Participants were permitted to assign any rating between 0 and 100 (inclusive) to the five-option multiple-choice items, and were instructed regarding the influence of guessing; however, all participants generated ratings in multiples of five.

Each of the items used in the practice session was accompanied by classical item analysis information (i.e., p-values, point-biserial correlations). Both groups were also given a questionnaire to collect background variables and to assess their perceptions of the adequacy of training in the methodology, comprehension of the standard-setting methodology,

perceptions of the ease of implementing the method, and confidence that the method would result in acceptably accurate separation of minimally-competent/not minimally-competent examinees.

After the practice ratings were completed, all members of both groups were polled to determine their perceived familiarity and comfort in proceeding with application of the method on operational test form. Following minor questions and clarifications, participants assigned to the group-process condition were given a copy of the operational test form, a test key, and a form for recording item ratings. This group remained in a group setting for the remainder of the meeting time. The group was encouraged to utilize their collective expertise and their packets of informational materials as needed.

A facilitator remained with the group-process condition group to monitor the discussion of items in that group, and to observe the frequency of discussion, the content of discussion, and the extent to which discussion was dominated by one or more group members. However, consensus for item ratings was not required, nor was any reviewer encouraged to change an item rating.

Participants assigned to the independent condition were separated from the whole group at the conclusion of the practice session. They were with the same booklet of test items, keys, and recording forms; however, they were instructed not to discuss their ratings with other members of the independent group, members of the group-process group, or other professional colleagues. The independent group provided their ratings independently and returned their completed rating forms by mail within two days of the whole-group meeting.

To address the second objective, item ratings members of the independent group were collected, duplicated, and returned to the group for a second round of ratings. For this

round, members of the independent condition group could see all ratings provided by each member for each of the items. With this normative information, this group then provided a second set of ratings, also independently and returned their ratings through the mail.

To address the third objective, information was obtained from travel industry sources relative to transportation, lodging, and per diem meal costs.

Results

Objective 1: Group versus independent ratings

Of primary interest was whether exposure to the two conditions (i.e., the independent rating of items or the use of the group-process method) resulted in differing overall passing standards. Table 1 provides descriptive statistics comparing the ratings produced under the two conditions. Visual inspection of the individual reviewer means listed in Table 1 suggests some interesting observations. First, each condition apparently contains one or more outliers. For example, while the reviewer means and standard deviations for the independent condition appear to be fairly similar (High to Low range of means equals 11.00) the variability of Reviewer 5's ratings is quite large compared to the rest of the reviewers in the independent condition. Similarly, in the group-process condition, Reviewer 10 produced an overall mean rating that was substantially lower that the other group-process condition reviewers. Of note also is that the variability of Reviewer 6's ratings is somewhat greater that the other reviewers in group-process group, although still not as large as the variability exhibited by Reviewer 5. Reviewer 10, who produced the lowest overall rating, also produced some disproportionately low ratings within the group-process condition, as evidenced by a relatively large positive value for skewness (1.373).

Insert Table 1 about here.

As a first step in exploring possible differences between the two conditions, two variables were created (INDEPRATE and GROUPRATE) to represent the overall rating for each item within each condition. INDEPRATE represents the mean rating for each of the 200 items across reviewers in the independent condition. GROUPRATE represents the mean rating for each of the 200 items across reviewers in the group-process condition. This procedure resulted in 200 pairs of ratings (one pair for each item). A correlation between the 200 pairs of ratings provided under each condition (i.e., between INDEPRATE and GROUPRATE) was also calculated and found to be .71, which was significantly different from zero at $p < .001$. Substantively, the magnitude of the correlation indicates that the reviewers in the two conditions exhibited a moderate degree of agreement on the mean rating for each item.

Table 2 presents a matrix of correlations for independent and group-process ratings. Individual reviewers' ratings did not necessarily correlate more strongly with other within-condition reviewers' ratings; the highest and lowest correlations were observed between conditions. A significance test of the mean transformed correlation between conditions was nonsignificant. However, within conditions, correlations between all reviewers' ratings were strong, positive, and significantly different from zero ($p < .01$).

Insert Table 2 about here.

Significant differences were not observed for the mean item rating provided under the independent (M=55.33, S=11.78) and group conditions (M=51.01, S=12.81). However, it should be noted that these values represent a proposed passing score for each condition, expressed as a percentage. That is, application of the standard proposed by reviewers in the group-process condition would result in a passing percentage of approximately 51% compared to the approximately 55% correct standard that would result from application of the standard based on the independent ratings. This means that, in raw score units, the independent condition mean of 55.33% correct would require examinees to respond correctly to 111 items in order to pass the examination, whereas the passing standard suggested by the group-process condition (51.01%) would be 102 items correct.

It seems worthwhile at this point to comment on the issue of practical impact in the absence of statistical significance. It is regularly observed that statistically significant findings can be of little practical importance (see, for example, Glass & Hopkins, 1996, p. 269). In the present study, however, although the mean difference in condition passing standards was not statistically significant, a substantial impact on pass/fail classifications would result from the nine-point raw-score scale differences in the suggested passing standards, acknowledging that such a difference may well be attributable to random variation.

In fact, the extent of classification changes that would be seen if the independent and group-process standards were applied to the actual distribution of scores observed for this examination was explored. Application of the group-process condition standard (approximately 102 items correct) would have resulted in a passing rate of 93.0% and a corresponding failure rate of 7.0%. On the other hand, had the independent condition standard been applied (requiring approximately 111 items correct), the passing rate would

have been 85.8% and the failure rate (14.2%) would have nearly doubled compared to the independent condition failure rate.

To ascertain whether overall variability in item ratings for the two groups was homogeneous, an F-test was performed ($F=1.18$, $df=199,199$; $p>.05$). This finding suggests that the overall ratings were not more variable under either of the two conditions.

Separate ANOVAs [with n = 200 items (random) and n = 5 raters (random)] were performed to learn if the overall ratings of individual reviewers within a condition differed significantly from each other. Additionally, the results of the two ANOVAs were used to address the question of whether exposure to either the independent or group-process condition was related to the variability of individual reviewers' ratings. Results of this analysis are presented in Table 3. The ANOVAs reveal a significant effect for raters in both the group-process ($F_{4,796} = 143.12$, $p < .001$) and independent ($F_{4,796} = 17.17$, $p < .001$) conditions, indicating that raters within a condition do produce different ratings (passing standards). As would be expected, the effect of items was also significant in both the group and independent conditions.

Insert Table 3 about here.

Decision Consistency

The extent to which exposure to the group-process condition and exposure to the independent condition results in differing levels of classification consistency was also examined, using the indices of classification consistency, p_o and κ. The group process condition exhibited a slightly greater index of overall consistency than the independent

condition (p_o = .958 and .930, respectively) although the contribution of the examination itself to consistency of classification decisions was slightly reduced under the group-process condition as compared to the independent condition (κ = .647 and .681, respectively).

Reviewers' ratings from both group and independent conditions were compared to the item difficulty values obtained from actual administration of the examination. For these analyses modified p-values were used, calculating the p-values based only of the responses of examinees (n=217) whose total score was within two standard errors of the operational passing score. Reviewer's ratings were found to be only weakly related to the modified p-values. All correlations between reviewers' ratings and modified p-values were significantly different from zero at $p<.001$, but ranged in magnitude from a low of .31 to a high of .42. Overall condition item rating correlations with the p-values were only somewhat greater (group r=.53; independent r=.54).

Relationship of Ratings to Observed Item Statistics

Two indices were created to reflect the degree of agreement between reviewers' ratings and two criteria. The first variable, E, was created to reflect the extent of agreement between a reviewer's ratings and the modified p-values. The variable E can be conceptualized as an index of absolute error. The second variable, E', reflects the degree of agreement between a reviewer's ratings and the mean ratings provided by reviewers within a particular condition. A comparison of the values of E and E' (shown in Table 4) across conditions indicated that reviewers exhibited large errors of specification, although they were much better at estimating how other reviewers in their condition would rate items than they were at predicting how the hypothetical minimally-competent group would perform. When evaluating the overall performance of the two conditions, it appears that the independent

condition results in slightly improved accuracy of specification in both the absolute and relative sense.

Insert Table 4 about here.

Objective 2: Use of normative information

To assess the impact of additional information, a second round of ratings was generated by item reviewers in the independent condition. In this case, the additional information consisted of distributions of ratings provided by independent-condition participants for the first 100 of the 200 items previously rated during round one. For the second round of ratings, participants were again mailed all materials and instructed to complete the rating task independently; however, they were encouraged to use the normative information.

Table 5 provides descriptive statistics comparing the ratings produced under the two conditions: "no-information" and "with-information." The no-information condition is defined as the independent provision by reviewers of Angoff ratings for the 100 items. The with-information condition is defined as the independent provision of a second set ratings for the same 100 items by the same reviewers, who were subsequently provided with the distribution of ratings generated under the no-information condition. As shown in Table 5, overall ratings produced with knowledge of the other reviewers' ratings are somewhat greater and less variable. Again, as in the no-information condition, one reviewer's pattern of ratings diverged from the trend suggested by the rest of the group. Examination of Table 5

reveals that Reviewer 5's overall ratings decreased under the with-information condition compared to the other reviewers whose overall ratings increased. Reviewer 5's overall with-information rating was also quite different from the fairly uniform overall with-information ratings provided by the other reviewers.

Insert Table 5 about here.

Two variables were created (NOINFO and WITHINFO) to reflect each item's overall rating under the two conditions. NOINFO represents the initial overall rating provided by the reviewers for each of the 100 items. WITHINFO represents the second (with information) rating provided by the reviewers for the same items. In each case, NOINFO and WITHINFO were obtained by calculating the mean rating for each item across reviewers within the no-information and with-information conditions. This procedure resulted in 100 pairs of ratings (one for each item).

The overall means for each condition and other descriptive statistics are presented in Table 6. The magnitude of the correlation between no-information and with-information conditions indicates fairly strong intra-reviewer agreement between initial and subsequent item ratings.

Insert Table 6 about here.

It should again be noted that the overall condition means represent the proposed passing score, expressed as a percentage, that would result from each condition. For

example, the no-information condition would result in a passing percentage of approximately 54.9% compared to the 60.1% that would result if the passing standard were established using the with-information ratings. This means that the no-information standard of 54.9% would require examinees to respond correctly to approximately 110 items on a full test of 200 items in order to pass the examination. The passing standard suggested by reviewers in the with-information condition (60.1%) would translate into a passing score of approximately 120 items correct on a 200-item test. This 10 raw score unit difference in passing scores over a 200-item examination reflects a difference of practical importance.

To test whether the difference in overall condition means was statistically significant, a repeated measures analysis of variance (ANOVA) was conducted. The full model specified three factors: Items (n = 100); Raters (n = 5); and Conditions (replication) (n = 2). Results of the repeated measures ANOVA are presented in Table 7. Despite the practical impact noted earlier, the analysis failed to reveal a statistically significant difference between the two conditions. However, inspection of Table 7 shows an expected significant effect for items and a significant effect for raters. Clearly, the results indicate that both items and raters differ in ways that affect the overall passing score.

A test for homogeneity of variances with paired (dependent) observations was also performed. The result was not statistically significant, suggesting that overall item ratings were not more or less variable under either condition.

Insert Table 7 about here.

Relationship between With-Information and No-Information Ratings

As shown in Table 6, a statistically significant correlation between no-information ratings and with-information ratings was observed ($r_{Noinfo,\ Withinfo} = .890$, $p < .001$). Calculation of the rank order correlation coefficient yielded similar results ($r_{Noinforank,\ Withinforank} = .871$, $p < .001$). These results indicate that the no-information and with-information conditions provided item ratings that were highly similar; that is, there is strong intrareviewer agreement between original and subsequent item ratings.

An intercorrelation matrix of item reviewers' first and second ratings was also produced and is presented in Table 8. Visual inspection of Table 8 reveals that reviewers' first and second ratings (i.e., under no-information and with-information conditions) are generally moderately correlated, ranging from a high of .759 (for Reviewer 5) to a low of .485 (for Reviewer 4) with a mean of .673.

Two groups of correlations are enclosed by dashed lines in Table 8. The encircled values correspond to the correlations based only on no-information ratings (upper left) and those based only on with-information ratings (lower right). After transforming these correlations using Fisher's r to Z transformation, a mean correlation for each condition was computed and the two overall means were compared. As hypothesized, the average with-information correlation exceeded the average no-information correlation (.473 > .337); however, the difference between the two mean correlations was not statistically significant.

Insert Table 8 about here.

Decision Consistency

The extent to which exposure to the no-information and with-information conditions resulted in differing levels of classification consistency was also examined. Indices of classification consistency p_o and k were calculated for each condition using the passing scores suggested by each. Application of the no-information passing score would result in a higher overall index of classification consistency ($p_o = .934$), compared to the with-information condition index ($p_o = .898$). Accordingly, the contribution to classification of the examination itself to consistency of pass/fail classifications was greater under the with-information condition ($k = .706$) compared to the no-information condition ($k = .678$).

Relationship of Ratings to Observed Item Statistics

For item reviewers in the no-information (NOINFO) and with-information (WITHINFO) conditions, overall item ratings for the 100 items were compared to item difficulty indices resulting from the actual administration of the examination. Modified p-values (MODP) were again used, obtained by calculating each item's difficulty based only on the responses of examinees whose total score was within two standard errors of the passing score.

Correlations were calculated between the overall NOINFO and WITHINFO ratings and MODP. Correlations were also calculated between individual item reviewers' ratings and MODP. For both conditions, individual reviewers' ratings were found to be moderately related to MODP. Interestingly, the lowest correlation with MODP ($r = .197$) was observed for a reviewer in the with-information condition, while the highest correlation with MODP ($r = .505$) was observed for a reviewer in the no-information condition. Also, surprisingly, the no-information condition produced a higher (though non-significantly) overall correlation

with modified p-values (r = .590) than the with-information condition (r = .573).

The two indices created to reflect the degree of agreement between reviewers' ratings and certain criteria (E and E') were also calculated for each reviewer. Table 9 presents the obtained values of absolute error of specification (E) and relative error of specification (E') for the five reviewers under no-information and with-information conditions. Comparison of the values displayed in Table 9 indicates that, generally, absolute errors of specification are only slightly reduced through the provision of additional information. The mean absolute error of specification for the with-information condition (24.12) was quite close to the mean for the no-information condition (24.93). However, relative errors of specification were also sightly reduced under the with-information condition (M = 13.43) compared to the no-information condition (M = 14.81).

Insert Table 9 about here.

In evaluating the effect of the provision of additional information, it is again observed that individual item reviewers were more proficient at estimating the overall group rating for the items than they were at predicting how the hypothetical minimally-competent examinee group would perform.

Regression Analyses

In order to further evaluate the effect of providing additional information to item reviewers, five regression analyses were performed. A regression model was developed which reflects the hypothesis that an individual reviewer's second (i.e., with-information)

rating can be predicted by knowledge of his original (without-information) rating and with knowledge of the group's original mean rating (with the group mean calculated excluding the individual reviewer). These two ratings were used as the independent variables in the regression equations with the reviewer's revised (with-information) rating used as the dependent variable. Theoretically, the model assumes that reviewers' make their judgments about item ratings based upon their own procedure-related knowledge; that is, knowledge regarding the hypothetical minimally-competent examinee group and the difficulty of the items being rated. And, reviewers take into account information gleaned from other reviewers; in this case, from the distribution of reviewers' initial ratings that was provided for their use in the second round of ratings.

To assess the likelihood of such an effect, five regression analyses were conducted, one for each reviewer according to the procedure described above. Results of the analyses are presented in Table 10. Raw (non-standardized) multiple regression equations are presented in the table, along with the correlations between the two independent variables, the Multiple R, and R squared. In each case, the correlations between the independent variables are low to moderate, suggesting that the choice of independent variables does not pose a threat of multicollinearity. For each regression performed, analyses of plots of predicted values against residuals revealed no disconcerting patterns; plots were broadly scattered and all residuals had means at or near zero.

Insert Table 10 about here.

The hypothesized influence of additional information appeared to be evident in each of

the regression analyses. For every reviewer, values of b_1 and b_2 were tested for significant difference from zero; in all cases, the test statistics were significant at $p < .01$. Further, the moderately high values of Multiple R and (with the exception of Reviewer 4) the moderate values of R^2 suggest that the regression model has accounted for at least half of the variation in reviewers' ratings.

Objective 3: Cost-effectiveness evaluation

As Norcini, et al. (1987) and Lockwood, et al. (1986) have noted, factors other than psychometric concerns can influence decisions regarding the conduct of passing score studies. Specifically, the cost of empaneling item reviewers may be prohibitive in many cases. Because differing costs would likely be associated with implementation of the group-process or independent conditions, an examination of the relative costs for each condition was undertaken.

For the following cost analyses, several assumptions were made. First, it was assumed that a group of participants ($n = 10$) were to be empaneled to provide ratings for a 200-item examination. For analysis of the group-process condition, it was assumed that nine of the ten participants would incur air travel, lodging, and meal expenses in order to travel to the passing score study site and participate in a standard-setting procedure lasting two days.

Two variations of the independent rating condition were explored in the analysis, in addition to the group-process condition. For one variation (hereafter called the "without-meeting" condition), it was assumed that the panel of reviewers would be mailed informational materials explaining the passing score methodology, then the test items to be reviewed would be rated independently and returned by mail to a central site. The second

variation of the independent condition (hereafter called the "with-meeting" condition), assumes that reviewers would travel to a single site for a one-half day meeting in order to become familiar with the passing score methodology. Reviewers in this condition would then receive a booklet of test items to be rated, would return to their cities of origin, and would return their ratings by mail.

Table 11 presents a summary of cost comparisons for the group-process condition and the two variations of the independent condition. Costs estimated in Table 11 are based upon figures published in the Travel Weekly (1996) for 1995, the most recent year for which complete information was available. This publication provides costs associated with travel expense categories ranked by major city and also provides a national average of expenses. To determine costs for this study, the national averages for airfare and lodging were used.

Insert Table 11 about here.

It should be noted that expenses associated with travel and consultation by a staff psychometrician, testing organization representative, or other personnel have not been included in the following analyses. Because licensure and certification boards vary in the extent to which they utilize in-house psychometric services or contract with external consultants, it was decided to omit this variable cost from each condition presented. Also excluded because of wide variability are expenses for conference room rental and equipment rental for the group-process and with-meeting conditions. Like the area of psychometric services, organizations vary widely in the extent to which they utilize "home office" facilities or conduct meetings off site. It is recognized that the exclusion of these expenses probably

results in a downward bias in the overall cost estimates for the group-process and with-meeting conditions.

Two additional assumptions should be noted. First, because air travel costs are extremely variable, depending on the city of origin, destination, class of service, and time of week, the costs for air travel were estimated to be $520.00 per person using figures obtained from Travel Weekly (1996) for round-trip weekday travel to and from "Anywhere, U.S.A." These airfare rates are based on advanced bookings that are less than two weeks in length--the most common type according to the business travel industry. As a result, these estimates may represent an upper bound for air transportation costs.

Also, the with-meeting variation assumes that reviewers would not require overnight lodging in order to participate in the meeting. In situations where overnight lodging is required due to distance travelled, flight connections, etc., lodging costs would be incurred. Thus, the with-meeting cost listed in Table 11 represents a lower bound estimate for that variation.

Examination of Table 11 suggests, based upon total costs for each of the three conditions, that the "Without-meeting" condition is by far the least costly method of conducting a passing score study. Total estimated costs for the three conditions are: Group-process condition, $7,520.00; With-meeting condition, $5,720.00; and Without-meeting condition, $290.00.

While it is true that the without-meeting condition is the least costly way of conducting a standard setting procedure when only monetary expenditures are considered, there are certainly other factors that should be discussed. For example, it was observed that the group-process conditions and independent conditions resulted in substantial variations in

passing scores. This is certainly not a factor that should be ignored. Another factor beside monetary cost that should be considered is the cost in terms of time. For many professions, it is quite difficult to identify experts who would be willing to forego two days of personal time or time away from professional activities in order to participate in a passing score study. For this reason, the "Without-meeting" condition, which would not require set-aside meeting time, could be viewed as the most economical. However, it should be noted that no data were collected as a part of this study to address the psychometric properties of the without-meeting condition. For that reason, this alternative can only be evaluated in terms of its economic feasibility and no conclusions regarding the accuracy or variability of without-meeting procedure results can be offered.

In summary, it was observed that expected savings in terms of time and financial resources were observed for the "With-meeting" and "Without-meeting" variations of the independent condition when compared to the group-process condition. However, it should be further noted that any savings incurred under any method would result in trade-offs that should be considered when those responsible for standard setting actually select a procedure. Also, some investigation of actual results from a without-meeting standard setting study seems warranted before any statements regarding its propriety should be made.

Discussion

This study demonstrated fairly large, though nonsignificant differences in results obtained by group and independent reviewers using the same standard-setting method on identical test content. Although the social interaction hypothesis would predict the observed results, the failure to achieve statistical significance for group mean differences does not rule

out the observation of these results due to chance, given the small sample size. One the other hand, small sample sizes are often used in practice; if replicated, these results should provide useful practical guidance. Also, it is regularly observed that statistically significant findings can be of little practical importance and nonsignificant findings lack practical value. In this study, although the mean difference in condition passing standards was not statistically significant, a substantial effect on pass/fail classifications was observed. In this case, the independent condition mean of 55.33% correct would require examinees to respond correctly to 111 of the 200 items in order to pass the examination, whereas the passing standard suggested by the group condition would be only 102 items correct. In the second part of this study, the no-information standard of 54.9% would require examinees to respond correctly to approximately 110 items on a full test of 200 items in order to pass the examination. The passing standard suggested by reviewers in the with-information condition (60.1%) would translate into a passing score of approximately 120 items correct on a 200-item test. Regardless of the statistical significance of the difference between the two condition means, the 10 raw score unit difference in passing scores over a 200-item examination clearly reflects a difference of practical importance.

The regression analyses revealed that ratings generated under the with-information condition were higher than ratings generated by the same reviewers under the no-information condition. The analyses demonstrated that a reviewer's "with-information" rating could be fairly well predicted by knowledge of the reviewer's original rating and knowledge of the group mean. These findings support the recommendations of others regarding the provision of normative information. The information may have had the effect of communicating to reviewers a group "expectation" or conceptualization regarding minimal competence levels

which they used in generating their second set of ratings. Accordingly, reviewers whose ratings may have been extreme initially were induced to converge on the standard implied by the distributions of item ratings, making their subsequent ratings for individual items somewhat less variable.

Finally, the results of the cost-benefit evaluation should provide additional information for entities engaged in standard setting to consider as they plan passing score studies. Both independent conditions appear to be more economically feasible for smaller certification boards; may require a reduced time commitment for participants; and--if research demonstrates that effective training materials can be developed and used independently by participants--may avoid some of the undesirable effects of the group process context.

REFERENCES

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (pp. 508-600). Washington, DC: American Council on Education.

Curry, L. (1987, April). Group decision process in setting cut off scores. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Fitzpatrick, A. R. (1989). Social influences in standard-setting: The effects of social interaction on group judgments. Review of Educational Research, 59, 315-328.

Glass, G. V, & Hopkins, K. D. (1996). Statistical methods in education and psychology, third edition. Needham Heights, MA: Allyn and Bacon.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15, 277-290.

Jaeger, R. M. (1988). Use and effect of caution indices in detecting aberrant patterns of standard-setting judgments. Applied Measurement in Education, 1, 17-31.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (pp. 485-514). New York: Macmillan.

Meskauskas, J. A. (1986). Setting standards for credentialing examinations: An update. Evaluation and the Health Professions, 9, 187-203.

Mills, C. N. & Barr, J. E. (1983, April). A comparison of standard setting methods: Do the same judges establish the same standards with different methods? Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. Journal of Educational Measurement, 24, 56-64.

Travel Weekly. (1996). 2(13), 18-20.

Table 1

Descriptive Statistics for Independent and Group-Process Reviewers

Independent Condition				Group-Process Condition			
Reviewer	Mean	Standard Deviation	Skew	Reviewer	Mean	Standard Deviation	Skew
1	60.23	17.67	-.255	6	45.57	21.91	.646
2	49.23	17.96	.252	7	60.68	16.56	-.561
3	57.23	16.66	-.201	8	64.13	15.24	-.771
4	51.18	17.26	.019	9	50.55	17.61	.140
5	58.79	26.55	-.217	10	34.13	14.41	1.373
Means	55.33				51.01		

Table 2

Correlation Matrix of Ratings from Independent and Group-Process Condition Reviewers

	Group-Process Condition Reviewers					Independent Condition Reviewers				
	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
R1	---	.339	.365	.289	.264	.425	.336	.451	.283	.328
R2		---	.243	.258	.253	.331	.290	.247	.172	.292
R3			---	.346	.225	.398	.347	.413	.317	.443
R4				---	.421	.301	.306	.351	.233	.351
R5					---	.299	.345	.175	.353	.414
R6						---	.371	.337	.296	.446
R7							---	.270	.308	.340
R8								---	.275	.388
R9									---	.320
R10										---

Table 3

ANOVA Results for Independent and Group-Process Conditions

	Independent Condition			Group-Process Condition		
Source	Mean Square	df	F	Mean Square	df	F
Items	820.03	199	2.99*	693.64	199	3.42*
Raters	4701.19	4	17.17*	29016.19	4	143.12*
Residual	273.79	796		202.74	796	
Total	5795.01	999		29912.57	999	

* = p < .001

Table 4

Absolute and Relative Errors of Specification for Item Reviewers in Independent and Group-Process Conditions

Independent Condition			Group-Process Condition		
Reviewer	E	E′	Reviewer	E	E′
1	23.46	14.31	6	28.10	15.99
2	26.42	15.58	7	23.49	15.74
3	23.72	13.51	8	24.56	17.75
4	24.53	13.26	9	25.33	13.47
5	27.90	19.59	10	32.37	19.86
Mean	25.21	15.25		26.77	16.56
Standard Deviation	1.90	2.59		3.57	2.39

Table 5

Descriptive Statistics for No-Information and With-Information Reviewers Across 100 Items

	First Rating (No Information)			Second Rating (With Information)		
Reviewer	Mean	Standard Deviation	Skew	Mean	Standard Deviation	Skew
1	57.35	17.56	-.141	62.50	18.46	-.382
2	52.95	19.93	.015	62.64	18.45	-.784
3	55.75	17.02	-.053	61.99	21.77	-.288
4	51.80	18.11	-.092	59.75	14.64	-.450
5	56.64	25.70	-.076	53.50	22.59	-.024

Table 6

Descriptive Statistics for No-Information and With-Information Condition Passing Scores

	NOINFO	WITHINFO
Mean	54.90	60.08
Standard Deviation	13.38	14.31
r NOINFO, WITHINFO =	.890	(p<.001)
Mean Difference =	5.178	

Table 7

Repeated Measures ANOVA Results for No-Information and With-Information Conditions

Source	Sum of Squares	Mean Square	df	F
Between Subjects				
Raters	9526.25	2381.56	4	11.71*
Within Subjects				
Conditions	881.75	881.75	1	0.77 ns
Raters x Conditions	4557.75	1139.44	4	
Items	179325.73	1811.37	99	8.91*
Items x Raters	80505.22	203.30	396	.72
Items x Conditions	10940.29	110.51	99	.39 ns
I x R x C, e	111660.91	281.97	396	
Total	397398.00	397.80	999	

* = $p<.001$

Table 8

Intercorrelation Matrix of Ratings from No-Information and With-Information Condition Reviewers

	No-Information Reviewers (Initial Rating)					With-Information Reviewers (Second Rating)				
	R11	R21	R31	R41	R51	R12	R22	R32	R42	R52
R11	---	.389	.367	.249	.231	.650	.484	.306	.255	.273
R21		---	.305	.259	.342	.522	.749	.372	.474	.399
R31			---	.338	.294	.455	.449	.722	.251	.452
R41				---	.457	.276	.473	.296	.485	.559
R51					---	.323	.449	.436	.244	.759
R12						---	.554	.433	.321	.401
R22							---	.483	.516	.517
R32								---	.294	.519
R42									---	.320
R52										---

Table 9

Absolute and Relative Errors of Specification for Item Reviewers in No-Information and With-Information Conditions

	No-Information Condition		With-Information Condition	
Reviewer	E	E′	E	E′
1	23.52	14.09	22.48	12.98
2	26.27	14.76	22.89	10.99
3	23.95	13.24	24.95	14.36
4	25.94	13.77	25.84	12.78
5	24.99	18.17	24.46	16.04
Mean	24.93	14.81	24.12	13.43
Standard Deviation	1.20	1.96	1.41	1.89

Table 10

Regression Analyses for Individual Reviewers

Reviewer	Regression Equation	$r\ x_1,x_2$	Mult. R	R^2
1	y = 8.805 + .535(x1) + .424(x2) + e	.425	.715	.511
2	y = 5.745 + .526(x1) + .524(x2) + e	.461	.827	.683
3	y =-1.073 + .789(x1) + .349(x2) + e	.456	.750	.563
4	y =29.528 + .290(x1) + .273(x2) + e	.480	.537	.288
5	y =-7.161 + .537(x1) + .555(x2) + e	.476	.807	.652

Notes: x_1 = original rating for item i by reviewer j, and x_2 = group's original mean rating for item i computed with reviewer j excluded.

Table 11

Comparison of Costs for Conducting a Passing Score Study under Group-Process and Independent Conditions

	Group-Process Condition	Independent Conditions	
Expense Category		With Meeting	Without Meeting
Meeting Time	2 days/2 nights	1 day/0 nights	0 days/0 nights
Air Travel	$4680.00	$4680.00	n/a
Lodging*	1710.00	n/a**	n/a
Meals***	700.00	350.00	n/a
Transportation+	400.00	400.00	n/a
Informational Mailing++	30.00	30.00	30.00
Test Items Mailing+++	n/a	130.00	130.00
Test Items Return Mailing+++	n/a	130.00	130.00
TOTALS	$7520.00	$5720.00	$290.00

Notes:

* = based on $95.00 per night rate.
** = assumes travel to and from meeting site in one day.
*** = based on $35.00 per diem.
\+ = assumes $20.00 per person each way to and from meeting site.
++ = first-class mailing costs only.
+++ = secure-method mailing costs only.